ASSOCIATED FACTOR OF MORTALITY RATE AMONGST PATIENTS WITH AIDS AND HIV-TB CO-INFECTIONS USING ZERO INFLATED NEGATIVE BINOMIAL METHOD

by

MOHD ASRUL AFFENDI BIN ABDULLAH

Thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy

UNIVERSITI SAINS MALAYSIA

2014

Acknowledgement

First of all, I would like to praise and thank Allah S.W.T and His mercy in completing this thesis. In particular, I am indebted to my supervisor Prof. Dr. Syed Hatim Noor for his trust, guidance and most importantly for the sense of humour. He spent time to study and improve my work from time to time to implement this thesis. I am proud to have as my supervisor as he never tired of giving advice and guidance in various ways. May Allah.sw.t bless him for his contribution and his kindness.

I would like also to express my gratitude to the Dean of the School of Medical Sciences and university, the Ministry of Higher Education, the University of Tun Hussein Onn, Malaysia on all financial assistance, my gratitude is also extended to all the staff in Unit of Biostatistics and Research Methodology, postgraduate center staff and medical record unit Hospital Universiti Sains Malaysia (HUSM). I am also grateful to Dr. Aniza Abdul Aziz and State Health department, Kelantan for kindly providing the data for the thesis.

Last but not the least, I also wish to thank my parents and my siblings for a lot of love and help in terms of support and encouragement to continue my studies up to PhD level. I would express my thanks to friends whose understanding and support me go on with my work till the end. INSYA ALLAH.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

NBM Negative Binomial Mortality

ZINBM Zero Inflated Negative Binomial Mortality

AIDS Acquired Immune Deficiency Syndrome

HIV Human immunodeficiency virus

TB Tubercolosis

HUSM Hospital Univesiti Sains Malaysia

HCFA Health Care Financing Administration

JCAHO Joint Commission for Accreditation of Healthcare Organization

AIC Aikake Information Criteria

BIC Bayesian Information Criteria

List of Symbols

$\beta = beta$

$\alpha = alpha$

$\theta = theta$

$\delta = delta$

$\epsilon = epsilon$

$\gamma = gamma$

$\eta = eta$

$\lambda = lambda$

$\mu = mu$

$\pi = pi$

$\sigma = sigma$

$\tau = tau$

$\phi = phi$

$\chi = chi$

$\psi = psi$

$\omega = omega$

$\Gamma = Gamma$

$\Delta = Delta$

$\Theta = Theta$

$\Lambda = Lambda$

$\Pi = Pi$

$\Sigma = Sigma$

$\approx = approximation$

$\partial = partial$

$\int = integrate$

$\sum = sum$

$\exists = exists$

$\downarrow = downarrow$

$\leftarrow = leftarrow$

$\rightarrow = rightarrow$

FAKTOR YANG MEMPENGARUHI KADAR KEMATIAN KE ATAS PESAKIT AIDS DAN HIV-TB DENGAN MENGGUNAKAN KAEDAH MODEL REGRESI NEGATIF BINOMIAL SIFAR MELAMBUNG

Abstrak

Banyak set data mempunyai ciri-ciri sebagai data kiraan dengan jumlah yang terlalu banyak. Data sebegini wujud di dalam pelbagai bidang kajian seperti kajian perubatan, kesihatan awam, toksikologi, epidemologi, sosiologi, psikologi, kejuruteraan pertanian dan sebagainya. Apabila pembolehubah bersandar adalah daripada pembolehubah bukan nilai negative, kebiasaannya model regresi Poisson di gunakan untuk menerangkan hubungan antara pembolehubah bersandar dan tidak bersandar. Walau bagaimanapun, jika di lihat daripada model Poisson sifar kita telah menyarankan model Poisson regresi sifar melambung lebih sesuai daripada model regresi Poisson yang sedia ada. Satu masalah yang dihadapi di dalam data ini ialah model sedia ada seperti model Poisson dan model Binomial gagal untuk menjelaskan perubahan yang wujud di dalam data ini. Selalunya data menunjukkan tambahan *serakan* yang terlalu banyak. Satu lagi komplikasi dalam data ini adalah berbentuk pembahagian data yang kadang-kadang terlalu jarang. Dalam kiraan kes Poisson, apa yang berlaku adalah terlalu kecil sama seperti kes model regresi Binomial. Jadi ia memerlukan kaedah yang sah untuk mengatasi masalah ini. Oleh itu, terdapat peningkatan sifar apabila model regresi Negatif Binomial Sifar Melambung di cadangkan kerana wujudnya *serakan* dan adalah lebih sesuai daripada model regresi Negatif Binomial yang sedia ada. Dalam kajian ini, umur adalah subjek bagi kadar kematian berubah-ubah mengikut kategori yang di buat. Data analisis yang di gunakan adalah data pesakit AIDS dan HIV-TB untuk membandingkan antara model regresi Negative Binomial dan model regresi

Negatif Binomial Sifar Melambung yang mana lebih baik unuk menentukan kadar kematian. Pemilihan model terbaik adalah berdasarkan model yang memberikan nilai cerapan yang paling kecil. Oleh itu, selaras dengan objektif umum untuk membandingkan model mana yang terbaik serta mengenal pasti faktor yang mempengaruhi kadar kematian ke atas penyakit tersebut. Ia di ikuti dengan menentukan *serakan* data di dalam model tersebut. Akhir sekali, dapatan yang di perolehi daripada analisa ini, model regresi Negatif Binomial Sifar Melambung adalah model terbaik serta faktor yang mempengaruhi kadar kematian AIDS adalah terdiri daripada golongan lelaki dan dari warganegara Malaysia. Bagi penyakit HIV-TB pesakit dari golongan yang kuat merokok menyumbangkan kadar kematian yang paling ramai di Kelantan.

ASSOCIATED FACTOR OF MORTALITY RATE AMONGST PATIENTS WITH AIDS AND HIV-TB CO-INFECTIONS USING ZERO INFLATED NEGATIVE BINOMIAL METHOD

Abstract

Many data sets are characterized as count data with a preponderance of zeros. Data in the form of counts and proportions arise in many fields such as studies in medicine, public health, toxicology, epidemiology, sociology, psychology, engineering, agriculture and soon. When the dependent variable is a nonnegative count variable, a Poisson regression model is commonly used to explain the relationship between the outcome variable and a set of explanatory variables. However, if extra-zero Poisson counts are observed, it has been suggested that a zero-inflated Poisson regression model is more appropriate than the classical Poisson regression model. One frequently encountered problem in these data is that simple models such as the Poisson and the Binomial models failed to explain the variation that exists. Often, data exhibit extra-dispersion (over or under dispersion). Another complication in data in the form of counts and proportions is that they are sometimes too sparse, that is smaller values have greater tendency to occur. In the Poisson case counts that occur are generally small and in the binomial case the binomial denominators are often small. Therefore, valid procedures are needed to detect departures from the simple models. Hence, when a lot of extra zero exists, zero inflated Negative Binomial has been suggested when overdispersion is present. It is more appropriate than the classical Negative Binomial regression model. Hence, this thesis follows the general objective, that is to compare Zero-Inflated Negative Binomial and Negative Binomial in identifying associated factors. The specific objective is to fit a Zero-Inflated Negative Binomial death rate regression model for mortality rate among AIDS/HIV co-infection patients and to compare Zero-Inflated Negative Binomial death rate regression with Negative Binomial death rate, which is the best model when

a data existing zeroes values. It follows by to determine overdispersion in the model. Lastly, to investigate the potential confounding factors affecting mortality rate among disease mapping co-infection patients among HIV-TB and AIDS. In this thesis, mortality rate is a subject of interest as dependent variable according to age categories by years. The data are analyzed from AIDS patients and HIV-TB mortality cases for comparing between Negative Binomial mortality and Zero Inflated Negative Binomial Mortality (ZINBM) which is better. Beyond this substantive concern, the choice should be based on the model providing the closest fit between the observed and predicted values. Unfortunately, the literature presents anomalous findings in terms of model superiority. In addition, the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) values were used to compare the fit between models. The results suggested that the literature are not entirely anomalous. However, the accuracy of the findings depended on the proportion of zeros and the distribution for the non zeros. ZINBDR tend to be the superior model, than the negative binomial model. The findings suggested there should be consideration of the proportion of zeroes and the distribution for the nonzero when selecting a model to accommodate zero-inflated data.

Chapter 1

Introduction

This chapter introduces motivations behind this thesis and presents a discussion of the primary aims and applications which are the development of exible and general statistical models for zero inflated negative binomial instead of zero inflated model. The structure of subsequent chapters in the thesis is then clearly laid out.

1.1 Motivation

In many areas of interest such as economic fields, agriculture, epediomolgy, ecology the dependent or response variable of interest (y) is a nonegative integer or count which is guessed to explain or determine in terms of a set of covariates (x). Unequal the traditional regression model, the response variable is discrete with a distribution that places probability mass at nonegative integer values only. In term of regression models for count, such other limited or discrete dependent variable model as well as the logit and probit, are linear with many conditions and special features intimately linked to discreteness and nonlinearity.

Hence, related with count observation Poisson Regression is frequently used to analyze based on count data as well (Argesti.A, 1996; Stokes.M.E. et al., 2000; YESILOVA.A.

et al., 2010; Cameron and Trevedi, 1998). In the classical regression model, Poisson Regression (PR) model explains the relationship between the dependent variable(y) and based on count independent variable or covariates (x).

Link function between the linear equation of dependent variable in given with the logarithmic transformation (Long.J.S, 1997; Lambert.D, 1992; Cameron and Trevedi, 1998). Assumption of PR requires means and variance which are equal to each other. But, in application normally it is not often possible to accomplish this assumption. If conditional variance is greater than conditional mean, it is absence overdispersion while vice-verse, conditional variance is lower than mean it is existing underdispersion (YESILOVA.A. et al., 2010; Wang.P. et al., 1996).

Generally in data set, existing overdispersion is seen more often than underdispersion which rarely happens. Existing overdispersion in PR regression model it might be biased to parameter estimated.Thus, Negative Binomial Regression (NBR) is appropriate to replace the PR when absence the overdipersion (Jansakul.N. and Hinde.J.P, 2008; Ridout.M., Hinde.J. and Demetrio.B.C.G, 2001; Lawles.J.F, 1987).

In NBR model, the parameter estimated are converged by considering effect that contains from overdispersion. Basically count observation might have excessive zero than expected. In such case ZIP regression model is appropriate approach to analyze the dependent variable having much zero observation (Lambert.D, 1992; YESILOVA.A. et al., 2010; Ridout.M., Hinde.J. and Demetrio.B.C.G, 2001; Bohning.D, 1998; Chueng.Y.B, 2002; Lee.A.H et al., 2001). Zero Inflated Poisson (ZIP) assumes that the population consists of two different types of observations where by one of them is based on count data consisting Poisson distribution that can have zero value exists (Bohning.D, 1998; Yau.W.K. et al., 2003). In such cases, when ZIP existing overdispersion and highly accessing zero

such mentioned above, Zero Inflated Negative Binomial (ZINB) is alternative method that is used (YESILOVA.A. et al., 2010; Long.J.S, 1997; Jansakul.N. and Hinde.J.P, 2008; Rose.C.E et al., 2006). According the discrete model such Poisson, NB, ZIP, and ZINB

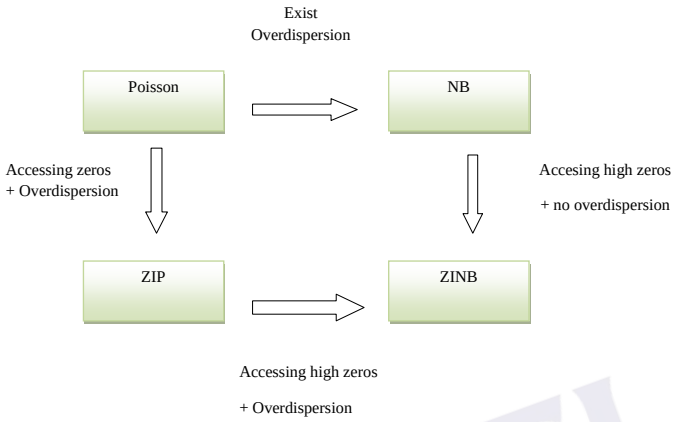

Figure 1.1: The frequently used models in the count data analysis framework

let us consider some examples from microeconomics, begining with samples independent cross-sectional observations, such ferteility study, frequent modeling number or live births over specified age interval of the mother, in which analyzing its variations in terms of mother schooling, age, household income, etc.

Other examples are, an accident analysis and model airline safety, where measuring the number of accidents experienced by an airline over some periods, and association between airline profitability and measures of the financial health of the airline (Rose.N, 1998; Winkelmann.R, 1990).Thus, seek to place a value on natural resources such as national forest by modeling the number of trips to a recreational site, into recreational demand of studies (Gurmu.S, 1977). Besides, model health demand studies based on the number of times that individual consumes a health services, such as number of visits to a doctor or days of stay in hospital in the past year and determine the impact of health status and health insurance (Cameron and Trevedi, 1998).

Meanwhile, in time series and panel data such using explanatory variables such as bank profitability, and bank borrowing from federal reserve bank . Analyzing a panel data example that has attracted much attention in the industrial organization literature on the benefit of research and development expenditure is the numbers of pattern received annually by firms. Refering to some cases as well , such as number of birth, the count is the variable of ultimate interest. Other cases, such as medical demand and results of research and development expenditure the variables interest are continous, often expenditure or receipts measured in dollar, but the best data available are instead a count (Hausman.J.A et al., 1990).

Hence, ZINB was used for predicting number of involved nodes in breast cancer patients (Dwivedi.A.K et al., 2010), which is fit and compared various count models to test model ability to predict the number of involved nodes. Subsequently, type I error rate also considers the use of count models for outcomes in randomized clinical trial setting instead of comparison model Poisson, Over-dispersed Poisson, Negative Binomial, ZIP, and ZINB. These methods in a series of simulation studies in application the ASAP (Addressing the Spectrum of Alcohol Problem)(Horton.N.J et al., 2007).

Parasites and vectors fields ZINB model is shown to be a useful tool for the analysis of individual based egg output data to measure is able to account for the disproportionately large number of zero eggs output. The probability of observing a zero egg count is demonstrated as being negatively associated with both female worm burden and male mean weight(Walker.M et al., 2009).

Discussing in health and quality of life outcome part of to investigate the association between oral health literacy (OHL) and oral health related quality of life (OHRQoL) and explore the racial differences there in among a low-income community based group of

female participants. The association of OHL with OHRQoL was examined using descriptive and visual methods, and was quantified using Spearmen's Rho and ZINB modeling (Divaris.K et al., 2011).

Subsequently, enviromental health ZINB was used to analyze protection from annual flooding is correlated with increased cholera prevalence in Bangladesh. How residence within or outside a flood protected area interacts with the probability of cholera prevalence (Carrel.M et al., 2010). Identify the mechanisms by which a neighborhood context affects aging will prepare us better for the coming decades, which will be characterized by a high urbanization rate and an increased number of people over to years of age. The component ZINB models that concern independent subjects predicts the probability of individuals belonging to group as compared with the group some different (Ferreira.F.R et al., 2009).

Negative binomial modelling in a longitudinal study of gastrointestinal parasit burdens in Canadian dairy cows. A ZINB model was applied to assess factors that would influence the fecal eggs counts with identified associations were eggs count were lowest in the winter and the highest in the late spring (Slymen.D.J.et al,2006).

However, in the psychiatry perspective ZINB was used to evaluate the relationship between depressed mood and receipt of mental health care services. In addition, this method examines the excess health care utilization due to job strain instead of both males and females have shown gender in differences of health care utilization(Gleicher.Y et al., 2011; Azagba.S and Sharaf.M.F, 2011).

In fact, ZINB models were run simultaneously to measure the likelihood of increased magnitude of disease events and the likelihood zero cholera or shigellolosis events (Car-

rel.M et al., 2011). Corresponding in this issues, ZINB made compared with other discrete count model in verbal fluency in Alzheimer's disease, Parkinson's disease and major depression part of assessing the sociodemographic and clinical factor associated with the disease severity (Araujo.N.B et al., 2011).

ZINB was used in the number of day prescriptions of Oral steroids in the year following date of statin initiation for the two exposure groups(Lodi.S et al., 2011). Meanwhile, to identify genomic regions enriched in a veriety of Chi-square and related next generation sequencing experiments (DNA-seq), calling both broad and narrow modes of enrichment across a range of signal to noise ratios. Hence, ZINB Alghorithm (ZINBA) was a proposed model and accounts for factor that co-vary with background and identifies enrichment in genomes with complex local copy number variations (Rashid.N.U et al., 2011).

Morever, ZINB models were used to predict hospitalization days and emergency room visits, including covariates of demoghrapic characteristic, employment status, psychiatric diagnosis, and concurrent substance use disoder. The main predictor variables of interest were receipt of illness management and recovery services, dropout from the program, and program graduation status (Salyers.M.P and Rollins.A.L, 2011). In journal of sex research, investigation the relationship between condom related protective behavioral strategies and condom use among college students ZINB were used to demonstrating that case (Lewis.M.A et al., 2011). According to HIV problem, ZINB method also compared part of modeling count outcomes from HIV risk reduction intervention, which is to analyze count outcome distributed with excess of zeros and overdispersion instead of appropriate fit models (Xia.Y et al., 2012).

1.2 Problem statement

Zero inflation distribution is a familiar statistical approach in a variety of decipline in the literature. Basically, from previous researchs, the results from both simulated and actual data sets involved the count distribution for dependent variable (y) and independent variable (x). Lambert (1992) found the that the ZIP model to be superior to the negative binomial ZIP model. When the data assesses higher zero value, overdisperison may exist. Hence, the negative binomial model and zero inflated negative binomial are appropriate to handle this situation.

One striking characteristic of these articles and others is their differences in terms of the proportion of zeros and the distribution for the non zeros. Bohning.et.al(1999) were analyzed data in which the proportion of zeros was as low as 0.216 and vice verse, Zorn (1996) used proprtions as high as 0.958. In fact, the nonzeros varied in terms of their distributions from highly positively skewed to normal to uniform. All these cases, the covariates is a count data.

Thus, in this research does it possible the negative binomial and zero inflation negative binomial working when the data acessing rate for dependent (y) and count values for covariates (x). This problem consider whereby the mortality rate among age categorical exists.Hence, to determine rate values can be working with independent count and zeros variables. Morever in Malaysia, the mortality rate of the disease maping such as AIDS/HIV, Tubercolosis, is a serious problem and should not be under wined. There are many reasons that lead to the occurrence of the problem. What are the factors that might be related or associated with the mortality rate among co- infection patients. There are several factors that are believed to affect mortality rate on this cases. Its may be used for the problem in term of mortality rate based on age categorical in other diseases.

Rate data such as rate of date during a week or a rate date per day. Subsequently, when an explanatory data are categorical, and occur in the same individual we simply need data entry process, whereby we make a calculation of the rate and do not distinguish between person- year follow up in that different individuals. It is possible that the different models yield different results depending on the proportion of zeros and the distribution for the nonzeros.

1.3 Rationale of the study

In various decipline studies as well, counts of a behavior or an event in a time interval of specified lenght are often collected observations, surveys, experiments or other statistical approach instead data observation. Hence, Cameron and Trivedi (1998) defined an event count as the number of times an event occurs, which is a realization of a nonnegative integer-valued random variable. Basically, count data contains excessive numbers of zeros value. In addition, these zeros can be categorized into two quantitative and qualitative zeros. A classical example involved quantitative and qualitative zeros such is counting the number of fish caught in a park during a period of time by multiple persons. Thus, some zeros results from fishing and not catching any fish (quantitaive zeros) and other zeros results from not fishing at all during that period of time (qualitative zeros). Subsequently, researchers have given this type count data as zero inflated count data.

Determination the best model is the one that appropriately answers the research question. Hence, superior model is one that has close proximity between the observed data and that predicted by the model. In other hands, a superior model is one with good fit to the data.

This study compared the fit between the negative binomial and zero inflated negative

binomial by age categorical death rate. Each analysis was performed for two different proportions of zeros value. The intended results would clarify the discrepant findings of previous research.

1.4 Objectives

The main objectives of this thesis is the development of the predicted statistical models for zero-inflated negative binomial mortality by age categories which may also have a general application.

1.4.1 General Objective

To compare between Zero-Inflated Negative Binomial mortality and Negative Binomial mortality and as well as to identify associated contributing factors.

1.4.2 Specific Objectives

(i) To fit a Zero-Inflated Negative Binomial mortality regression model and to investigate the association factors affecting mortality among AIDS co-infection patients.

(ii) To fit a Zero-Inflated Negative Binomial mortality regression model and to investigate the association factors affecting mortality among HIV-TB co-infection patients.

(iii) To compare Zero-Inflated Negative Binomial mortality regression and Negative Binomial mortality, which is the better model when a data existing zeroes values.

(iv) To determine overdispersion in the model.

1.5 Significance of the study

The primary purpose of this study was to determine superiority of fit for various models mortality disease mapping by categorical age death rate. As such, determination can be made as to which model has better fit given data with proportion of zeros and a particular distribution. The superior model is the appropriate model given the research question. Hence, there are situations in which the appropriate model is unknown or unclear. Further, there may be situations in which a simpler model such as the Binomial, Poisson may be used to subtitute of the more sophisticated zero inflated models. This research provides results that aid researchers in determining model to use given zero-inflated data.The significant of the study are;

- The main research might create a model for Zero-Inflated Negative Binomial Death Rate (ZINBDR) co-infection patients.
- This research is important to organization (hospital) to identify which possible explanatory variable can give effect on disease.
- This research can integrate knowledge for proposed new technique to solve a problem occur with medical statistic programming in order to create up to date decision tools for strategic and subsequent operational decision.

1.6 Research Question

This part is explaining a research study according the models distribution were used. Hence, model comparisons in this research were based on two measures. One is the deviance statistic which is measure of the difference in log-likelihood between two models, permitting a probabilistic decision as to whether one model is adequate or whether an alternative model is superior. This statistic is appropriate when one model is nested within another model. The other measure is Akaike's Information Criteria (AIC) and Bayesian

Information Criteria (BIC).

These two measures of model fit were used to compare results from where each data set included 945 cases HIV/AIDs and 176 data set cases HIV+TB co-infections patients. Specifically, the measures of model fit were used to answer the following research question;

- Let given one two-level categorical covariate with known values and continous covariate with known values, what is difference in the estimated log-likelihood between the negative binomial vs. zero inflated negative binomial death rate?.
- Assume one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated AIC and BIC between all the models?.

Chapter 2

Literature Review

One refers to the particular issues with negative binomial while the other relates to the zero inflation of methodology. The aim in this chapter is to review associated key literature and to establish general concepts relevant to subsequent chapters.

2.1 Zeros Model

Group 1 with probability α and is in group 2 with probability $1-\alpha$. Here, α is an unknown parameter that is to be estimated. The first group consists the subjects who always have zero counts. For example, a scientist who will never publish, perharps because of the nature of his/her job, would be in this group. On the other hand, a scientist with zero publications is in the first or the second group. If we did, this could be entered explicitly in the regression as an independent variable. Hence, the distinction between the two group is a form of discrete, unobserved heterogeneity (Long.J.S, 1997)

Hence, the second group counts are governed by a PRM or NBRM. Such the Poisson case;

$$Pr(y_i|x_i) = \frac{\exp(-\theta_i)\theta_i^{y_i}}{y!}$$

where $\theta = \exp(x\beta)$. Zero counts occur by chance with probability $Pr(y = 0|x) = \exp(-\theta)$. This corresponds to the scientist who tries but fails to publish. Zero counts are generated by two different processes, depending on the group. The overall probabilities of *0's* from each group, weighted by the probability of an individual being in that group. According the equation below;

$$Pr(y_i = 0|x_i) = [\alpha \times 1] + [(1 - \alpha) \times \exp(-\theta)] = \alpha + (1 - \alpha)\exp(-\theta_i)$$

Since the Poisson process only applies to $1 - \alpha$ of the sample, the probability of positive counts must be adjusted as follows;

$$Pr(y_i|x_i) = (1 - \alpha)\frac{\exp(-\theta_i)\theta_i^{y_i}}{y!};\ y > 0$$

(Prove that $\sum Pr(y|x) = 1$).

2.2 Zero Inflated Models

Zero-inflated distribution originated from the work of Rider (1961) and Cohen (1963) which is the first concept who examined the characteristics of mixed Poisson distribution. Mixed Poisson distribution are characterized by data that have been mixed with two Poisson distributions in the proportions α and $1-\alpha$, respectively. Let λ_1 and λ_2 be the parameters of two Poisson distribution and the probability function of the mixed distribution as follows;

$$f(x) = \alpha\frac{\lambda_1^x e^{-\lambda_1}}{x!} + (1 - \alpha)\frac{\lambda_2^x e^{-\lambda_2}}{x!}, x = 0, 1, 2, ...,$$

where without any loss of generality, $\lambda_2 > \lambda_1$ (the means of the two distribution) and n is the observed count data $(0, 1, 2, \ldots, n)$. The k^{th} factorial moment of x may be written as;

$$m_{(x)} = \alpha\lambda_1^k + (1-\alpha)\lambda_2^k = \alpha(\lambda_1^k - \lambda_2^k) + \lambda_2^k$$

Both Rider (1961) and Cohen (1963) have proposed different approaches using the method of moments for estimating the parameter α. Cohen further described an approach for estimating the parameter α with zero sample frequency. Johnson and Kotz (1969) were the first to explicitly define a modified Poisson distribution (known as Poisson with added zeros) that explicitly accounted for excess zeros in the data. The modified distribution is the following;

$$\begin{aligned} f(x) &= \alpha + (1-\alpha)e^{-\lambda}; n = 0 \\ f(x) &= (1-\alpha)\frac{e^{\lambda}\lambda^n}{n!}; n \geq 1 \end{aligned}$$

Thus, Johnson and Kotz (1969) proposed a similar procedure to the one suggested by Cohen (1963) for estimating the parameter α. Under this distribution, $n = 0, 1, 2, \ldots, K$ are inflated counts while the rest of the distribution $K+1, K+2, \ldots, N$ follows a Poisson process.

The concept of the mixed Poisson distribution introduced by the previous authors has been particularly useful to describe data characterized with a preponderence of zeros. For this type of data, more zeros are observed that would have been predicted by a normal Poisson or Poisson-gamma process. It is generally believed that data with excess zeros come from two sources or two distinct distributions, hence the apply-named dual -state process. The underlying assumption for this system is that excess zeros solely explain the heterogeneity found in the data and each observation has the same mean λ.

Two different types of regression or predictive models have been proposed in the literatures for handaling this types of data. The first type is known as the hurdle model. The zero-inflated count models (also called zero-altered probability or count models with added zeros) represent an alternative way to handle data with a preponderance of zeros. Since their formal introduction by Lambert (1992), they use of these models has grown almost boundlessly and can be found in numerous fields, such traffic safety, economics, epidemiology, sociology, trip distribution, and political science among others.

2.3 Zero Inflated Distribution

Zero-inflated distribution can be explained as a distribution, when there are more zeroes than would be expected for a typical distribution. This can be modeled as a mixture of two distribution which is one degenerate at zero with probability P and the other one some count, with probability $(1 - P)$. The general probability model of mixture zero such as;

$$P(Y = y) = \begin{cases} p + (1-p)g(0) & \text{for} \quad y = 0 \\ \\ (1-p)g(y) & \text{for} \quad y_i > 0 \end{cases} \tag{2.1}$$

where $g()$ is discrete distribution function. Zero inflated distributed has been used to model count or abundance data from examples such as working by Lambert (1992). All previous literature there is a large probability at zero but relatively predictable probabilities elsewhere. The model for a ZINB as folows, where P_{zeroes} is the probabilty of the zero

class,$x = 0, 1, 2, \ldots, n$ for $0 < p < 1$;

$$\begin{aligned} P(X = 0) &= P_{zeroes} + (1 - P_{zeroes})(n-1)p^n \\ P(X = x) &= (1 - P_{zeroes})\binom{n+x-1}{x}p^n(1-p)^x \end{aligned}$$

The frequency of zero counts in the previous literature was much larger than frequency of any other counts. The model fitted the data very well, and the authors suggested choosing a model based on dispersion of the non-zero counts. Hence, (Warton, 2005) fit the data existing a large probabilities at zero with transformed distribution, log-linear distribution, negative binomial distribution using method of moment estimation (MME) and maximum likelihood estimation (MLE) ZIP and ZINB. Thus, the large zero counts in abudance data were move likley to have arisen from a negative binomial distribution with a small mean than from a ZINB distribution. In addition, (Welsh.A.H et al., 1996) suggested testing whether the zero inflation term was necessary and importantly that NB distribution generally abudance modeling data better than other distributions even when there were more zeros than predicted by the models.

2.4 Zero Inflated Data

As a count data model, that name implies a data from counting process. Thus, part of rate basically implies that when event of a certain type occur overtime, space, or some other index of size. Futhermore, the model might describe how the rate depends on explanatory variables instead of the response values take from of discrete integers(Zorn.C, 1996).

Subsequently, Cohen (1963) concerns over zero-inflation without covariates for count data, while (Cameron and Trevedi, 1998) had identified many areas in which special model to analyze count data such as bank failure, occupational injuries and illness, number of patients and so on. In many application, most of dissertation had been done on frequency of

event whereby focus on count data. In this dissertation, we use dependent variable (y) as rate by age categorical death rate of AIDS and HIV- TB in Kelantan area, Malaysia.

An example of rate data variables such as city unemployment rate, median income, and percentage of residents having completed high school. Thus, modeling rates can model a proportion with logistic regression and it allow for time at risk (exposure). According a rate data exposure often measure in person-years whereby model a rate based on incidents per unit time.

1. There is a rate at which events occur.
2. This rate may depend on covariate
3. Rate must be more than 0.
4. Event are independent
5. Then the number of event observed will follow a discrete distribution.

2.4.1 Zero-Inflation

Normally it is not uncommon for the outcome variable count data distribution to be characterized by propondence of zero. Basically for a count data the outcome variable measure an amount that must be non negative and may in some causes be zero. The positive values are generally skewed, often extremly and the distribution of data of this types follow a common from; there is a spike of discrete problem mass at zero, followed by a bump describing positive values"(Tooze.J and Jores.R,2002).

To model a cases according a rate function, basically used a count function as a guideline to subtitute be a rate. This is a primary such in the case of internal or ratio count data. Hall and Berenhaut (2002) explained that based on countinous data these distribution have

a null probability of yeilding a zero.

There is a little motivation for a model such as zero inflation normal, because all observed zeros are unambiguous. Subsequently, they can be analyzed separately from the non zeros, if continuoses zero are inflated. The null probability of continuous zeros is evident in measures such as height and age.

Meanwhile, many author have a own ideas to simulate about zero-inflation.The condition of excess zero is known as zero inflation and gives as a probability mass that clumps at zero. Hence, Min and Agresti (2004) formally defines zero inflation as *"data for which a generalized linear model has lack of fit due to disprotionately many zero.There are simply."*

2.4.2 The sources of zero-inflation

Like a count data, in this case the zeros can be classified being either the zeros or sampling zeros. For a true zeros that shown responses of zeros that are truly null. Let a student attend workshop in the college,*"How many students prepared for attending a workshop?"*. In this situation, some of the respondents in the sample have no attention to share about it. Thus, the number of students attended a workshop may never be greather than zero.

Hence, sampling zero on the other hand, arise as a probability. There are proportion of student who have not attended a workshop due to possibility that the workshop was not or is not yet availaible. Other way, some student may fell proposed and have no reason to participate in a workshop.

The mechanism underlying zero-inflation can arise from one or both of;

- A possibility that no other response is probabilistic or;

- That the response is within a random sample of potential count or rate response.

The term of sampling zeros as *"false zeros"* and included error as a sources of zeros. Again,they mention Min and Agresti (2005) ,*"zero inflation is often the result of a large number of 'true zero'observation. The term zero inflation can also be applied to data set with 'false zero'observation because of sampling or observer errors in the data collection."*

2.4.3 Impact of zero-inflation on analysis

Zero inflation is a mostly often in count data modelling. It appears from the recognition that the use of continous distribution to model integer outcomes might have unwelcome consequences including inconsistent parameter estimates. However, in a count data scenerio, the zero left bound implies heteroscedasticity(Zorn.C, 1996).

While (Tooze.J,K and Jones.R.,2002) the large problem with zero-inflation distribution beyond this inadequately the analyzing such a skewed and heteroscedastic distribution as if it were normal and that they yield suprisingly large inefficiency and nonsensicial results (King.G, 1989). Beside,(G et al., 2005; McCullagh.P. and Nelder.J.A, 1989) explain that the zero inflation is a special case of overdispersion in which the variance is greather than mean it should be given a particular distributional shape and measure of central tendency. The impact is biased or inconsistent particular estimates, inflated standard errors and invalid inferences. To solve a problem according zero-inflated might such as follows;

(a) *Deleting zeros*

Regarding (Tooze.J,K and Jones.R.,2002) deleted all cases having responses of zero on the variable of interest is the simplest solution to solve zero inflated. However, a large proportion of total responses would then the removed from a total data set. This results in a loss of value able information impacting statistical conclusion va-

lidity. To analyze sample size, it may also then be too small for analysis of the non-zero values.

(b) *Transforming zero*

In count data problem, to control a normal distribution is called transforming the counting process (Slymen.et.al,2006). Based on count distribution, its often appear to be positively skewed, one reasonable transformation involves taking the natural logarithm of the responses to the predictor variables. Beside (King.G, 1989; Zhou.X and Tu.W, 1999) they assuming the zeros have not been deleted, the transformation will not work since the natural logarithm of zero is undefined. While, since transformation is linear, this technique has been shown to yield parameter estimates that differ as a function of the adjustment quantity (King.G, 1989). The undefined log zero problem has been handle, the original problems (Welsh.A.H et al., 1996) had state, it is clear for data with many zeros values that such on approach will not be valid as the underlying distributional assumption such as linearly and homocedasticity. It is be voilated. Lastly in any techniques of problem, transformation sometimes create a new problem while solving the old one a transform that produces constant variance may not produce normality (Argesti.A, 1996)(p.73).

2.4.4 Example of zero-inflated data

$Example$ 1: Self-reported counts of specific "high-risk" behaviors in a given time period.Heilbron (1994) examined the data on counts of a "high-risk" heterosexual behavior (anal intercourse) which grossly suggested the addition of zero counts. The data are from the U.S.A. National AIDS Behavioral Study, for a subset of "Center City" respondents aged 18 - 49 years old who reported having had heterosexual or bisexual relations during the last 5 years, and having 2 - 12 sexual partners in the past 12 months, but not reporting some other risk factors of HIV infection (hemophilia, injection drug use in the last 5 years, positive HIV antibody test). The response analyzed was the reported count of times the

respondent had anal intercourse with partners of the opposite sex during the last 6 months. Two independent variables were considered: gender; and whether or not the respondent had a "risky" main sexual partner in the last year. "Risky" indicates presence of any of the risk factors that were excluded in respondents. Among 1244 qualifying respondents in the sample, 129 excluded as having missing values on one or more of the three variables. Table 2.2 presents the distribution of the response, within subsets defined by the two independent variables. Examination of the fit from log-linear GLM's confirmed the presence of added zeros.

Table 2.1: Distribution of the response

	Subset number gender risk factor			
	1	2	3	4
Y	Male	Male	Female	Female
0	541	102	238	103
1	19	5	8	4
2	17	8	2	2
3	16	2	1	.
4	3	1	1	1
5	6	4	1	.
6	5	1	1	.
7	2	.	.	.
8	6	.	.	.
9	1	.	1	.
10	.	1	.	.
12	3	.	.	.
15	1	.	.	.
20	.	1	.	.
30	.	.	.	.
37	.	.	.	.
50	.	.	.	1
	620	125	253	117

2.5 Review of prototype of the models

In this section, I will reconsideration several related statistical models and theories that will be subsequently extended. Such the Poisson regression model for count data; the zero-inflated Poisson regression model for zero -inflated count data; beta regression model for

fractional data observed on [0,1),(0,1] or [0,1] ; and zero inflated beta regression models.

2.5.1 Poisson Regression Model

We consider now the Poisson regression model (PRM). The Poisson regression model is the most basic model. With this model the probability of a count is determined by a Poisson distribution, where the mean of the distribution is a function of the independent variables. This model has the defining characteristic that the conditional mean of the outcome is equal to the conditional variance. The Poisson regression model is an example of GLM in which the distribution of the response Y with covariate vector x is Poisson with density such equation below. Let y be a random variable indicating the number of times that an event has occured during an interval time, y has a Poisson distribution with parameter $\theta > 0$ if (Argesti.A,1996);

$$Pr(y|\theta) = \frac{e^{-\theta}\theta^{y}}{y!}$$

As θ increases, the mass of the distribution shifts to the right as follows;

$$E(y) = \theta$$

the parameter θ is known as the rate since it is the expected number of times that an event has occured per unit of time. θ can also be thought of as the mean or expected count. The variance equals the mean such;

$$Var(y) = E(y) = \theta$$

the equality of the mean and the variance is known as $equidispersion$. The PRM is a special case of the generalized linear model with exponential family distribution (McCullagh.P. and Nelder.J.A, 1989) and is also a special case of a nonlinear regression

model. The basic assumption such all events are independent of each other and all events have a constant arrival rate through the fixed parameter β. Thus, one implication of the model specification is that both the conditional mean $E[y_i|x_i]$ and the conditional variance $V[y_i|x_i]$ are equivalent to $\exp(x_i'\beta)$ due to the statistical properties of the Poisson distribution. Hence, count variables often have a variance greater than the mean, which is called $overdispersion$. Dealing with over-dispersion or under-dispersion problem, negative binomial model can be used (Cameron and Trevedi, 1998) because it assumes the conditional distribution follows a negative binomial distribution.

The development of many models for count data is an attempt to account for overdispersion. As θ increases, the probability of 0's decreases, such as for θ=0.6, the probability of a 0 is 0.55; for θ=1.5,it is 0.22; for θ=3.0, it is 0.05, and for θ=10.5, the probability is 0.00002. Thus, there are more observed 0's than predicted by Poisson distribution instead of count variables.

Poisson distribution can be derived from a simple stochastic process, known as Poisson process, where the outcome is the number of times that something has happened. A critical assumption of a Poisson process is that events are $independent$ which is an event occurs it does not affect the probability of the event occuring in the future.

2.5.2 Zero Inflated Poisson

For handling data with excess zeros, Lambert (1992) introduced the zero-inflated Poisson (ZIP) model. The ZIP model uses a mixture link function approach with two link functions, a logit link function and a log link function, to capture the statistical features of the two processes: perfect zero state and Poisson process. Combining the Poisson count

model and the binary process for the ZIP model as follows;

$$\begin{aligned} Pr(y_i = 0|x_i) &= \omega_i + (1-\omega_i)\exp(-\theta_i); \\ Pr(y_i|x_i) &= (1-\omega_i)\frac{\exp(-\theta_i)\theta_i^{y_i}}{y!}; y_i \geq 1 \end{aligned}$$

Thus shows that;

$$Var(y_i|x_i, z_i) = [0 \times \omega_i] + [\theta_i \times (1-\omega_i)] = \theta_i - \theta_i\omega_i$$

The conditional mean of the model has been changed by lowering the expected count by $\theta\omega$. The conditional variance of ZIP model such as;

$$Var(y_i|x_i, z_i) = \theta_i(1-\omega_i)(1+\theta_i\omega_i)$$

Subsequently, in this model the logit link is used to predict the conditional probability, ω of a subject i to be in the Poisson process. The log link is used to predict the expected mean θ of event counts for subject i given that the subject is not in the perfect zero state, but rather in the Poisson process. Thus, β and γ are the associated regression parameters.

As the negative binomial model handles the overdispersion or under dispersion problem for regular count data, zero inflated negative binomial model (ZINB) with a dispersion parameter, θ can handle both excess zeros and dispersion problems.

2.5.3 Beta Regression Model

Beta distribution is very flexible and its occur in proportion, fraction and rates modelling. Its commonly used regression models to analyse that are perceived to be related to other variables. Subsequently, the authors had extended beta distribution according the problem was occurred. Its, desire to investigate how certain variables of a continous variable

Bibliography

Albalak.R, Kamemerer.S and O'Brein.R.J (2007), 'Trend in tubercolosis/humman immunodeficiency virus comorbidity, united states 1993-2004', *Arch Intern Med* pp. 2443–2452.

Anderson.R.N. and Rosenberg.H.M (1998), Age standardization of death rates: Implemntation of the year 2000 standard, Technical Report 3, National Center of statistics.

Anscombe.F.J (1950), 'Sampling theory of the negative binomial and logarithmic series distributions', *Biometrika* pp. 358–382.

Araujo.N.B, Barca.M.L, Engedel.K and Coutinho.E.S.V (2011), 'Verbal fluency in alzheimer's disease, parkinson's disease, and major depression', *Journal Clinical Science* pp. 623–627.

Argesti.A (1996), *An introduction to categorical data analysis*, second edn, John Wiley and Son.,Hoboken, New Jersey.

Azagba.S and Sharaf.M.F (2011), 'Psychosocial working conditions and the utilization of health care services', *BMC Public Health* pp. 1–7.

Bains.N (2009), Standardization of rates, Technical report, Associtaion of public health epidemologists in Ontario(APHEO).

Bates.MN, Khalakdina.A and Pai.M (2007), 'Risk of tubercolosis from exposure to tobacco smoke: a systematic review and meta analysis', *Plos Med* pp. 335–342.

Blom.G (1958), *Statistical Estimates and Transformed Beta Variables*, Wiley,New York.

Bohning.D (1998), 'Zero inflated poisson models and c.a.man. a toturial collection of evidence', *Biometrical JournaL* pp. 833–834.

Böhning.D, Dietz.E, Schlattmann.P, Mendonça.L and Kirchner.P (1999), 'The zeroinflated poisson model and the decayed, missing and filled teeth index in dental epidemiology', *Journal of the Royal Statistical Association, Series A, 162* pp. 195–209.

Cameron and Trevedi (1998), *Regression Analysis for Count Data*, Cambridge University Press,New York.

Carrel.M, Escamilla.V, Messina.J and Winston.J (2011), 'Diarrheal disease risk in rural bangladesh decreases as tubewell density increases: a zero-inflated and geographically weighted analysis', *International Journal of Health Geographics* pp. 1–9.

Carrel.M, Voss.P, Streatfields.P.K, Y.Mohammad and Emch.M (2010), 'Protection from annual flooding is correlated with increased cholera prevalence in bangladesh: a zero-inflated regression analysis', *Journal Environmental Health* pp. 1–9.

Choi.B.C.K, deGuia.N.A and Walsh.P (1995), 'Look before you leep:stratify before you standardize', *American Journal of Epidemiology* pp. 1087–1095.

Chueng.Y.B (2002), 'A study of growth and development', *Statistic in Medicine* pp. 1461–1469.

Clarke (2001), 'Testing nonnested models of international relations: Reevaluating realism', *American Journal of Political Science* pp. 724–744.

Cohen.A.C (1963), 'Estimation in mixture of discrete distribution : In proceeding of the international symposium on discrete distribution,montreal,quebec'.

Dean.C and Lawless.J.L (1989), 'Test for detecting over-dispersion in poisson regression models', *Journal of the American Statistical Association* pp. 467–472.

Divaris.K, Lee.J.Y, Baker.A.D and Jr.W.F (2011), 'The relationship of oral health literacy with oral health-related quality of life in a multi-racial sample of low-income female caregivers', *Journal Health and Quality of Life Outcomes* pp. 1–9.

Dwivedi.A.K, Deo.S, S.Rakesh and K.Elizabeth (2010), 'Statistical model for predicting number of involved nodes in breast cancer patients', *National Public Access* pp. 641–651.

Ezzati.M and Murray.M (2007), 'Tobbaco and tubercolosis: a quantative systematic review and meta-analysis', *Plos Medical* pp. 2206–2216.

Farewell.V.T and Sprott.D.A (1993), 'The use of a mixture models with count data', *Statistic and probability letter* pp. 53–60.

Ferreira.F.R, Cesar.C.C and Camargos.V.P (2009), 'Aging and urbanization: The neighborhood perception and functional performance of elderly persons in belo horizonte metropolitan area-brazil', *Journal of Urban Health: Bulletin of the New York Academy of Medicine* pp. 54–65.

G, M., A, W., Rhodes.J.R, Kunnert.P, A, F., J, L.-C., J, T. and Possingham.P (2005), 'Zero tolerance ecology: Improving ecological inference by modeling the source of zero observations', *Ecology Letters* pp. 1235–1246.

Gajalakshmi.V, Peto.R, Kanaka.T.S and Jha.P (2003), 'Smoking and mortality from tuberculosis and other diseases in india: retrospective study of 43,000 adult male deaths and 35,000 controls', *Am J Respir Crit Care Med* pp. 507–515.

Gleicher.Y, Croxford.R, Hochman.J and Hawker.G (2011), 'A prospective study of mental health care for comorbid depressed mood in older adults with painful osteoarthritis', *BMC Psychiatry* pp. 1–10.

Gurmu.S (1977), 'Test for detecting overdispersion in the positive poisson regression model', *Journal of Business and Economic Statistics* **9**(2), 215–222.

Guthrie.JR and Smith.AM (1995), 'Physical activity and the menopause experience: a cross sectional study in maturitas 1995', *Arch Intern Med* pp. 71–80.

Hall.D and Berenhaut.K.S (2002), 'Score test for heterogeneity and overdispersion in zero-inflated poisson and binomial regression modelsl', *The Canadian Journal of Statistics* **30**(3), 1–15.

Hall.D.B (2003), 'Zero-inflated poisson and binomial regression with random effect: A case study', *International for Quality in Health Care* **4**(15), 319–329.

Hausman.J.A, Hall.B.H and Griliches. (1990), 'Econometric model for count data with an application to the patents-r and d relationship', *Econometrica* pp. 909–938.

Henn.L, Nagel.F and Pizzol, D. (1999), 'Comparison between humman immunedeficiency virus positive and virus negative patients with tubercolosis in southern brazil', *Mem Inst Oswaldo Cruz* pp. 377–81.

Holmes.C.B, .H, H. and Nunn.P (1998), 'A review of sex of sex diferences in the epidemology of tubercolosis', *Internationak Journal Tuberc Lung Disease* pp. 96–104.

Horton.N.J, Kim.E and Saitz.R (2007), 'A cautionary note regarding count models of alcohol consumption in randomized controlled trials', *BMC Medical Research Methodology* pp. 1–9.

Jang.T.Y (2005), 'Count data models for trip generation', *Journal of Transportation Engineering* pp. 444–450.

Jansakul.N. and Hinde.J.P (2008), 'Score test for extra zero model in zero-inflated negative binomial models', *Computation in Statistics-Simulation and Computation* **38**(1), 92–108.

Joseph.M.Hilbe (2011), *Negative Binomial Regression*, Cambridge University Press,New York.

Kahn.K.L, R.H, B. and Draper.D (1988), 'Interpreting hospital mortality data:how can we proceed?', *JAMA* pp. 260–3625.

Kibria.B.M.G (2006), 'Application of some discrete regression models for count data', *Pakistan Journal Statistical Operation Research* **1**, 1–16.

King.G (1989), 'Event count models for international relations: Generalizations and applications', *International Studies Quarterly* pp. 123–147.

Lambert.D (1992), 'Zero-inflated poisson regression with an application to detects in manufacturing', *Technometrics* **34**, 1–14.

Landon.B, Iezzoni.L and Ash.A.S (1996), 'Judging hospitals by severity adjusted mortality rates: The case of cabg surgery', *Inquiry* pp. 155–166.

Lawles.J.F (1987), 'Negative binomial and mixed poisson regression', *The Canadian Journal of Statistic* pp. 1–13.

Lee.A.H, Wang.K and Yau.K.K.W (2001), 'Analysis of zero inflated poisson data incorporating extent of exposuret', *Biometrical Journal* pp. 963–975.

Lewis.M.A, Kaysen.D.L, Rees.M and Woods.B.A (2011), 'The relationship between condom-related protective behavioral startegies and condom use among college students: Global and event level evaluations', *Journal of Sex Research* pp. 471–478.

Lindquist.O and Bengstsson.C (1980), 'Menopausal age in relation smoking', *Arch Intern Med* pp. 147–149.

Lodi.S, Carpenter.J, Egger.P and Evan.S (2011), 'Design ofcohort studies in chronis diseases using routinely collected databases when a prescription is used as surrogate outcome', *Medical Research Methodology* pp. 1–9.

Long.J.S (1997), *Regression models for categorical and limited dependent variables*, Sage publication.

Lundberg.O (1993), 'The impact of childhood living conditions on illness and mortality in adulthood', *Social Science Medicine* pp. 1047–1052.

Lyles.R.H., Lin.H.M. and Williamson.J.M (2006), 'A practical approach to computing power for generalized linear model with norminal, count, or ordinal responses', *Statistics in medicine* **26**, 1632–1648.

Matsui.S (2005), 'Sample size calculation for comparative clinical trials with over-dispersed poisson process data', *Statistic in Medicine* **24**, 1339–1356.

Mazerolle.M.J (2004), 'Mouvements et reproduction des amphibiens et tourbiès perturbèes'.

McCullagh.P. and Nelder.J.A (1989), *Generalized Linear Models*, 2nd edition edn, Chapman and Hall.

Min.Y and Agresti.A (2004), Random effects models for repeated measures of zeroinflated count data, Technical report, Department of Statistics, University of Florida.

Ospina.R and Ferrari.L.P (2010), 'Inflated beta distribution', *Journal Applied Statistic* pp. 111–126.

Park.R.E and Kosecoff.J (1990), 'Explaining variations in hospital death rates: Randomness, severity of illness, quality of care', *JAMA* pp. 264–484.

Peck.MN (1994), 'The importance of chilhood socio-economic group for adult health', *Social Science Medicine* pp. 553–562.

Penner.C, Roberts.D, Kunimoto.D, Manfreda.J and Long.R (1995), 'Tuberculosis as a primary cause of respiratory failure requiring mechanical ventilation', *Am J Respir Crit Care Med* pp. 867–872.

Rashid.N.U, Giresi.P.G and Ibrahim.J.G (2011), 'Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, event within amplified genomic regions', *Genome Biology* pp. 1–20.

Rico.P (1999), 'Who, health situation analysis and trend summary', *Demography* pp. 296–315.

Rider.P.R (1961), 'Estimation the parameter of mixed poisson, binomial and weibull distribution by method of moments', *Bulletin de I'Institute de statistique 38, part 2* .

Ridout.M., Hinde.J. and Demetrio.B.C.G (2001), 'A score test for testing a zero-inflated poisson regression model againts zero-inflated negative binomial alternatives', *Biometrics* **57**(1), 219–223.

Ridout.M., Hinde.J. and Demetrio.G.B (2001), 'A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives', *Biometric* **57**, 219–223.

Rose.C.E, Martin.S.W, Wannemuehler.K.A and Plikaytis (2006), 'On the zero-inflated and hurdle models for modeling vaccine adverse event count data', *Journal of Biopharmaceutical Statistical* pp. 463–481.

Rose.N (1998), 'Economic determinants of airline safety performance', *Journal of Political Economy* pp. 944–964.

Salem.S.S, Moore.G, Rucker.M and Pearson.S (1994), 'The case for case-mix adjustment in practice profiling', *JAMA* pp. 871–874.

Salyers.M.P and Rollins.A.L (2011), 'Impact of illness management and recovery programs on hospital and emergency room use by medicaid enrollees', *NIH Public Access* pp. 509–515.

Sanguanwongse.N, Cain.KP and Suriya.P (2008), 'Antiretrovviral therapy for hiv infected tubercolosis patients saves lives but needs to be used more frequently in thailand', *Journal Acquir Immune Deficiencies Syndrom* pp. 81–89.

Self.S. and Mauritsen.R.H (1991), 'Power/sample size calculation for generalized linear models', *Biometrics* **44**(373), 79–86.

Shieh.G (2001), 'Sample size calculation for logistic and poisson regression models', *Biometrica* **88**(4), 1193–1199.

Shieh.G (2005), 'On power and sample size calculation for wald test in generalized linear models', *Journal of Statistical planning and inference* **128**, 43–59.

Signoroni.D.F (1981), 'Sample size for poisson regression', *Biometrica* **78**(2), 446–450.

Simas.B.A and Rocha.V.A (2010), 'Improved estimator for a general class of beta regression models', *Computional Statistic and Data Analysis* pp. 348–366.

Slymen.D.J, Ayala.G.X, Arredondo.E.M and P, E. (2006), 'A demonstration of modeling count data with an application to physical activity', *Epidemiologic Perspectives and Innovations* pp. 1–9.

Stokes.M.E., Davis.C.S. and Koch.G.G (2000), *Catogerical data analysis using the SAS system*, 2nd edn, John Wiley and Son.

Vuong.Q.H (1989), 'Likelihood ratio tests for model selection and non-nested hypotheses', *Econometrica* pp. 307–333.

Walker.M, Hall.A, Anderson.R.M and Basanez.G (2009), 'Density dependent effects on the weight of female ascaris lumbricoides infections of humans and its impact on patterns of egg production', *Parasit and Vectors* pp. 1–18.

Wang.P., Puterman.M.L., Cockburn.L. and Le.N (1996), 'Mixed poisson regression models with covariates dependent rates', *Biometrics* **52**(2), 381–400.

Warton (2005), 'Many zeroes does not mean zero inflation: Comparing the goodness-of-fit of parametric models to multivariate abundance data', *Environmetrics* pp. 275–289.

Web.GB (1998), 'The effect of the inhalation of cigarette smoke on the lungs. a clinical study', *Am Journal Tubercolosis* pp. 25–27.

Welsh.A.H, Cunningham.R.B, Donnelly.C.F and Lindenmayer.D.B (1996), 'Modelling the abundance of rare species: Statistical models for counts with extra zeroes', *Ecological Modelling* pp. 297–308.

Whittemore.A.S (1981), 'Sample size for logistic regression with a small response probability', *Journal of the American Statistical Association* **76**(373), 27–32.

Williamson.J.M., Lin.H.M., Lyles.R.H. and Hightower.A.W (2007), 'Power calculation for zip and zinb models', *Journal of Data Science* pp. 519–534.

Winkelmann.R (1990), 'Duration dependence and dispersion in count data models', *Journal of Business and Economic Statistic* pp. 467–474.

Xia.Y, Beedy.D.M, Ma.J and Feng.C (2012), 'Modeling count outcomes from hiv risk reduction interventations: A comparison of competing statistical models for count responses', *AIDs Research and Treatmenat* pp. 1–11.

Yau.K.K.W and Lee.A.H (2001), 'Zero inflated poission regression with random effect to evaluate an occupational injury prevention programme', *Statistic in Medicine* **20**, 2907–2920.

Yau.W.K., Wang.K. and Lee.A.H (2003), 'Zero-inflated negative binomial mixedregression modelling of overdispersed count data with extra zeros', *Biometrical Journal* **45**(4), 437–452.

YESILOVA.A., KAYA.Y., KAKI.B. and KASAP.I (2010), 'Analysis of plant protection

studies with excess zeros using zero-inflated and negative binomial hurdles mode;', *Gazi University journal science* **23**(2), 131–136.

Zhou.X and Tu.W (1999), 'Comparison of several independent population means when their samples contain log-normal and possibly zero observations', *Biometrics* pp. 645–651.

Zorn.C (1996), 'Evaluating zero-inflated and hurdle poisson specifications', *Midwest Political Science Association* pp. 1–16.